MISAGGREGATION EXPLAINS CONSERVATIVE INFERENCE ABOUT NORMALLY DISTRIBUTED POPULATIONS

Gloria E. Wheeler, et al

University of Southern California

Prepared for:

Office of Naval Research
Advanced Research Projects Agency

1 Aug 1975

001597-5-T

UNIVERSITY OF SOUTHERN CALIFORNIA

social science research institute

TECHNICAL REPORT

MISAGGREGATION EXPLAINS CONSERVATIVE INFERENCE ABOUT NORMALLY DISTRIBUTED POPULATIONS

GLORIA E. WHEELER

AND

WARD EDWARDS

SPONSORED BY:

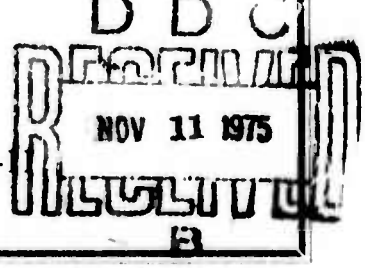
ADVANCED RESEARCH PROJECTS AGENCY
DEPARTMENT OF DEFENSE

MONITORED BY:

ENGINEERING PSYCHOLOGY PROGRAMS
OFFICE OF NAVAL RESEARCH
CONTRACT No. N00014-75-C-0487, ARPA
ORDER #2105

AUGUST 1975

SSRI RESEARCH REPORT 75-



Social Science Research Institute
University of Southern California
Los Angeles, California 90007
213-746-6955

The Social Science Research Institute of the University of Southern California was founded on July 1, 1972 to permit USC scientists to bring their scientific and technological skills to bear on social and public policy problems. Its staff members include faculty and graduate students from many of the Departments and Schools of the University.

SSRI's research activities, supported in part from University funds and in part by various sponsors range from extremely basic to relatively applied. Most SSRI projects mix both kinds of goals — that is, they contribute to fundamental knowledge in the field of a social problem, and in doing so help to cope with that problem. Typically, SSRI programs are interdisciplinary, drawing not only on its own staff but on the talents of others within the USC community. Each continuing program is composed of several projects; these change from time to time depending on staff and sponsor interest.

At present (Spring, 1975), SSRI has four programs:

Criminal justice and juvenile delinquency. Typical projects include studies of the effect of diversion on recidivism among Los Angeles area juvenile delinquents, and evaluation of the effects of decriminalization of status offenders.

Decision analysis and social program evaluation. Typical projects include study of elicitation methods for continuous probability distributions and development of an evaluation technology for California Coastal Commission decision-making.

Program for data research. A typical project is examination of small-area crime statistics for planning and evaluation of innovations in California crime prevention programs.

Models for social phenomena. Typical projects include differential-equation models of international relations transactions and models of population flows.

SSRI anticipates continuing these four programs and adding new staff and new programs from time to time. For further information publications, etc., write or phone the Director, Professor Ward Edwards, at the address given above.

MISAGGREGATION EXPLAINS CONSERVATIVE INFERENCE ABOUT NORMALLY DISTRIBUTED POPULATIONS

Technical Report
1 August 1975

Gloria E. Wheeler

and

Ward Edwards

Social Science Research Institute
University of Southern California

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Engineering Psychology Programs, Office of Naval Research under Contract No. N00014-75-C-0487, ARPA, Order #2105.

Approved for Public Release
Distribution Unlimited

SSRI Research Report 75-11

10

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 001597-5-T	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Misaggregation Explains Conservative Inference About Normally Distributed Populations	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER: None
7. AUTHOR(s): Gloria E. Wheeler and Ward Edwards	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0487
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Social Science Research Institute, University of Southern California, Los Angeles, California 90007	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: ARPA Order No. 2105
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Boulevard, Arlington, Virginia 22209	12. REPORT DATE: 1 August 1975
	13. NUMBER OF PAGES: 41
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Engineering Psychology Programs, Office of Naval Research, Arlington, Virginia 22217	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Conservatism
Probabilistic Information Processing
Likelihood Ratio
Posterior Odds
Bayes Theorem
Misperception
Misaggregation
Response Bias

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Three major hypotheses have been proposed to account for conservative inference: misaggregation, misperception, and response bias. The research reported in this paper allowed the testing of these hypotheses. Subjects made probabilistic judgments about stimuli generated from normally distributed populations. The populations were piles of pick-up sticks, each stick having one end painted blue and the remainder painted yellow. The length of blue paint was the random variable. In Experiment 1, each S made 4 types of

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102 LF 014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(Cont.)

judgments: noncumulative likelihood ratios, noncumulative odds, cumulative likelihood ratios, and cumulative odds. The results indicated that there was little difference between likelihood ratio and odds judgments, and that when judging single stimuli, Ss were veridical; conservatism only occurred when Ss were in a cumulating condition. Thus the results ruled out misperception hypothesis.

Experiments 2 and 3 varied d', sequence construction, and population display. Sequences were constructed that would accentuate differences between predictions made by response bias and misaggregation hypotheses. The data showed that subjects made veridical independent trial estimates but aggregated information conservatively, regardless of how far odds and likelihood ratios were from 1:1, thus permitting rejection of most forms of the response bias hypothesis.

Misaggregation Explains Conservative Inference About Normally Distributed Populations

Gloria E. Wheeler and Ward Edwards
Social Science Research Institute
University of Southern California

A standard finding in the literature about probabilistic inference is that people are conservative; that is, probabilistic data cause less change of opinion than is appropriate. For reviews of the literature, see DuCharme (1969), Edwards (1968), or Slovic and Lichtenstein (1971). Try it yourself. Think of two bookbags filled with poker chips (traditional apparatus for such experiments). Bag R contains 70% red chips and 30% blue ones; Bag B contains 30% red and 70% blue. You flip a fair coin to choose a bag without knowing which it is. You sample randomly 12 times from the chosen bag, replacing the chip after each sample, and get 8 red chips and 4 blue ones. Write an estimate on the margin of this page of the probability that you sampled from Bag R. Most people estimate in the range from .60 to .85. The correct probability is .967.

Bayes's Theorem of probability, the appropriate normative model for such probabilistic inferences, may be written

$$\Omega_1 = L_1 \Omega_0 \qquad (1)$$

$$\log \Omega_1 = \log \Omega_0 + \log L_1 \qquad (2)$$

Ω_0 is the prior odds; $\Omega_0 = \frac{P(H_A)}{P(H_B)}$.

L_1 is the likelihood ratio appropriate to the datum D_1; $L_1 = \frac{P(D_1|H_A)}{P(D_1|H_B)}$.

Ω_1 is the posterior odds, or odds appropriate after the datum has been observed; $\Omega_1 = \frac{P(H_A|D_1)}{P(H_B|D_1)}$. If Equation (2) (say) is used to process D_1 and then datum D_2 comes along, the appropriate additional processing of course is

$$\log \Omega_2 = \log \Omega_1 + \log L_2$$

$$= \log \Omega_0 + \log L_1 + \log L_2$$

$$= \log \Omega_0 + \log (L_1 L_2)$$

Note that the likelihood ratio appropriate for several data (here assumed to be conditionally independent of one another) is the product of the likelihood ratios for each datum considered separately. This multiplicative combining rule for likelihood ratios is the same as the multiplicative combining rule by which likelihood ratios combine with prior odds to specify posterior odds; both are direct consequences of the product rule for probabilities of independent events.

These formal rules imply a list of steps required to perform probabilistic inference properly. Perhaps people perform intuitive analogs of each step when performing such tasks. What step or steps go seriously wrong? For a critical review of possibilities and data bearing on them, see Edwards (1968).

The misperception hypothesis asserts that people incorrectly perceive the diagnostic impact of each datum; they in effect estimate L_1 and L_2 as closer to 1:1 than they should. The misaggregation hypothesis asserts that people may perceive the diagnostic impact of any single datum correctly (i.e., may correctly estimate Ls) but fail to aggregate data properly, either with other data or with priors. The response bias hypothesis can take various forms; a typical one simply asserts that as evidence piles up and odds or probabilities get extreme, subjects become progressively more reluctant to use such extreme numbers. Obviously, these three classes of hypotheses are not mutually exclusive; moreover each is a class, not a single hypothesis. Still they have been the main contenders.

Various experiments bear on these classes of hypotheses. For example, Edwards, Phillips, Hays, and Goodman (1968), in a complex simulation experiment found that disaggregated likelihood ratio estimates lead to considerably more extreme posterior odds than do aggregated posterior odds or posterior probability estimates. They interpreted their findings as evidence for misaggregation, but the evidence is equivocal because their situation precluded calculation of normatively correct posterior odds; directly estimated posterior odds could have been correct and those calculated from likelihood ratio estimates could have

been extreme. Peterson, DuCharme, and Edwards (1968) and Wheeler and Beach (1968) found that subjects interpreting binomial data produced conservative multi-datum likelihood function estimates. Such evidence, however, only shows that whatever produces conservatism in posterior estimates based on multiple data also produces conservatism in multiple-data likelihood function estimates; it does not bear on misaggregation vs. misperception vs. response bias.

Binomial data are unsatisfactory for such experiments, since only two single-datum likelihood ratios are possible on a given trial. Two normal distributions differing only in mean are better, since a datum may produce any likelihood ratio. DuCharme and Peterson (1968) used the heights of men and women and found little conservatism, possibly because subjects are so familiar with the stimuli. They used only four-item sequences, so their subjects did relatively little aggregating. They displayed their two normal distributions in several ways, all orderly.

The present experiments also use two normal distributions differing only in mean, but the materials are abstract and non-numerical (painted pick-up sticks) and the displays of distributions are disorderly. Most important, the sequences are up to ten items long. This permits comparison of single-stimulus estimates with cumulative estimates requiring considerable mental aggregation.

Experiment 1

Methods

Subjects. Thirty-six naive male university students, run individually or in pairs, were the subjects.

Apparatus. The stimuli for the experiment were 7" long pick-up sticks (from the well-known children's game). One end was painted blue; the other, yellow. The point along the stick at which blue switched to yellow varied for different sticks. The length of blue (or, because of the symmetry of the situation, the length of yellow) was the random variable. The two populations of sticks were normally distributed with means of 4.50" and 2.50" respectively, and a common standard deviation of 1.25". Single stimuli were shown to subjects vertically in a white wooden holder, constructed so that the stick was held at its ends and all 7" could be seen. For several-stick sequences, the sticks were leaned against a white background in the order in which they had been sampled.

The population characteristics were displayed for the subjects by two random histograms, each representing a random sample of about 100 sticks from one

of the populations, arrayed left-to-right in the order of sampling. The displays were actual size and colors, and on each the population mean was displayed by a heavy black horizontal line at the appropriate position. These displays were visible throughout the experiment. The lengths displayed had in fact been carefully chosen to represent the populations accurately.

Responses were made in 10-page booklets, one response per page. On each page was printed:

_____:1 in favor of hypothesis _____.

Sequences. Eight sequences of 10 normal deviates each were drawn from a table of random normal deviates. Some were slightly modified to produce a fairly large range of final posterior odds and to eliminate unduly long runs of rare stimuli, etc. Sixteen physically different sets of 10 pick-up sticks were prepared from these eight mathematical sequences; each normal deviate was represented once as more blue than yellow and once as more yellow than blue. Each stick sequence was used twice, so each subject saw 32 sequences of sticks. A single random order of sequences was used for all subjects.

For this experiment, $d' = (m_1 - m_2)/\sigma = 1.6$. The likelihood ratio at the mean (of either population) is 3.60. So typical veridical odds might be about 365,600:1 after 10 sticks. The smallest 10-stick posterior odds was 52,520; the largest was 877,820.

Response modes. Each subject made responses in four response modes: noncumulative likelihood ratio, cumulative likelihood ratio, noncumulative odds, and cumulative odds. The 36 subjects were split into four groups which used the response modes in different orders. For example, Group 1 responded to each stick in the first 8 sequences with a noncumulative likelihood ratio, to each stick in the second 8 sequences with a cumulative likelihood ratio, and so on. No effect of response mode order was found, and so data for a given response mode are here aggregated over response mode order.

Instructions. The experimenter described the charachteristics of the pick-up sticks and pointed out that the charts on the wall were representative of the kind of sticks one would expect to get in drawing randomly from each population. Next the experimenter read the appropriate response mode instruction for the first phase of the experiment, and proceeded with 8 sequences of sticks. After

that, the experimenter read the next appropriate response mode instruction, proceeded with 8 more sequences and so on. Subjects were told not to look back on their previous responses at any time.

The four response modes were explained as follows.

Noncumulative Likelihood Ratio. The experimenter explained that she would show the subject a stick and he was to determine which of the two piles would be more likely to produce a stick like that one and how much more likely it was. He was told that a likelihood ratio of 1:1 meant that either pile was just as likely to produce that stick and that a likelihood ratio of 10:1 meant that there were 10 times as many sticks like that in the favored pile as in the unfavored pile.

Cumulative Likelihood Ratio. If the subject had already done the noncumulative likelihoood ratio task, the experimenter indicated that during this phase of the experiment he would also estimate likelihood ratios. If he had not performed the noncumulative task, the experimenter first explained it as above. In either case, the experimenter then explained that first the subject would see and evaluate a single stick. Then the experimenter would display a second stick and the subject was to evaluate the likelihood ratio of both sticks. He was told to forget that he had seen the first stick by itself and assume that the sticks had been presented simultaneously. He was to determine which pile was more likely to produce that sample of sticks, and how much more likely it was. Both sticks were displayed at this time. After the subject made his estimate for the sample of two sticks, the experimenter displayed a third stick and the subject estimated the likelihood ratio for the sample of three. He was instructed that this procedure would continue until 10 sticks were displayed, and then it would start over again.

Noncumulative Odds. The subject was asked to assume that the experimenter had flipped a fair coin to select pile A or B, then had drawn one stick at random from the chosen pile. His task was to determine which pile the experimenter was drawing from and to estimate the odds in favor of that pile. To give him an understanding of odds he was hold that an estimate of 1:1 would mean that he was completely uncertain about which population was being sampled; odds of 10:1 meant that if he were to see that stick 100 times, about 91% of the time it would have come from the more likely pile. After the subject made his judgment,

the experimenter again selected one population by a flip of a coin and drew a single stick randomly from it.

Cumulative Odds. The subject was told to assume that the experimenter had flipped a fair coin to select one of the populations, then had drawn 10 sticks one at a time from the chosen pile. The subject would see the sticks in the order that the experimenter had drawn them. Upon seeing the first stick, the subject was to make an odds estimate. (If the subject had not already done the noncumulative odds part of the experiment, he received an explanation about what an odds estimate was.) After seeing the second stick, he was to revise his odds to take into account the new information in the second stick. Then he would see and judge a third stick, a fourth stick, and so on, until he had seen all 10 sticks that had been drawn from the pile. After all 10 sticks had been displayed, a pile was again selected randomly and the experimenter drew 10 sticks from the newly chosen pile.

Results and Discussion

First, the data were subjected to a logarithmic transformation, and all analyses were performed on the log transformed responses. Moreover, the correct hypothesis is represented in the numerator of every odds and every likelihood ratio. Log odds or log likelihood ratios less than 0 nevertheless occasionally appear, since the first stick or two may favor the incorrect hypothesis.

In the analysis to follow, the dependent variable is described as "mean inferred log odds". In Bayes's Theorem [Equation (1)], the posterior odds is obtained by multiplying prior odds by the likelihood ratio appropriate to the datum. Because in this experiment the prior odds at the beginning of each sequence was 1:1, the numerical value of the cumulative log likelihood ratio is equal to the value of the log posterior odds. In the noncumulative response modes, one can infer the log posterior odds for any sequence by applying Bayes's Theorem to the subjects' responses; that is by adding the single-stick log likelihood ratios or the single-stick log posterior odds (which are numerically though not conceptually equal to log likelihood ratios). Thus the phrase "mean inferred log odds" refers to posterior odds obtained either directly from the subjects' mean log responses (for the two cumulative response modes) or by applying Bayes's Theorem to their mean log responses (for the two noncumulative response modes).

Scatterplots showing mean inferred log odds on the ordinate and the Bayesian log odds on the abscissa appear in Figure 1. All four correlation coeffi-

---

Insert Figure 1 about here

---

cients are greater than .90, indicating considerable orderliness in the mean data. But the slope of the best fitting regression line is not the same for all response modes. The noncumulative response modes, whether likelihood ratio or odds, show near-veridicality, with slopes of 1.04 and 1.17 respectively. The two cumulative response modes show conservatism; the slopes for likelihood ratios and odds are .28 and .36 respectively. These findings show little inherent difference between likelihood ratios and odds. Evidently subjects perceive the impact of single items of information rather accurately, in the sense that they can accurately estimate single-stick odds or likelihood ratios. The fact that, for either likelihood ratios or odds, conservatism only occurs when subjects must cumulate a sequence of data is strong evidence against misperception as the explanation of conservatism.

Further evidence against the misperception hypothesis is obtained by looking at regression analyses of Bayesian log likelihood ratios vs. mean inferred log likelihood ratios as a function of sequence trial number. Inferred likelihood ratio is taken as the number the subject estimated for the noncumulative response modes. For cumulative response modes, it is obtained for the Nth datum by dividing the response made to it by the response made to its predecessor. Figure 2 shows that, for all four response modes, the slope of Trial 1 is nearly 1.0, indicating little or no conservatism. It stays near 1.0 on Trial 2 for all

---

Insert Figure 2 about here

---

response modes, although the cumulative likelihood ratio responses show some conservatism. By Trial 3, both cumulative response modes show considerable conservatism, which further increases on Trial 4. In fact, for all trials beyond Trial 2, both cumulative response modes show strong conservatism. Neither of the noncumulative response modes show conservatism, nor do they show any effect of trial number. Since every noncumulative response was essentially a Trial 1

response, one would not expect to find a relationship between trial number and slope of the regression line for those two response modes.

These data clearly refute the misperception hypothesis. Subjects consistently did very well at estimating single-stick numbers. They also show little difference between odds and likelihood ratio estimates--a convenience for designers of experiments and of information processing systems.

Veridical responses in cumulative response modes were very large indeed--usually well over 100,000:1 by the 10th stick. Are subjects unwilling to estimate large numbers, and is that the reason for conservatism? The question deserves an experiment of its own.

Experiment 2

DuCharme and Peterson (1968) ran an experiment using normal data generating processes in which the two populations under consideration were men and women, and the random variable was the height of an individual. The experimenter would select one of the two populations and sample heights randomly from it; the subject's task was to guess which population the experimenter was sampling from and to give an odds estimate in favor of that population. DuCharme and Peterson found very little conservatism, and that which they did find tended to favor the response bias hypothesis.

DuCharme (1970) then ran a series of two experiments further investigating the subject's responses to normal data generating processes. The first experiment differed only slightly from the DuCharme and Peterson (1968) experiment except that in the second part of it the prior odds for selecting one of the populations were different from 1:1. In the second experiment he used several different pairs of normal populations (fictitious species of fish). The length of the fish was the random variable. His results strongly supported the response bias hypothesis, although not in its most simple form. He argued that within the odds range of 1:1 to about 10:1, people will generally be veridical. When the correct values fall outside that region, people will become less and less accurate, since they have had very little experience with large numbers (such as 100,000:1) and those numbers are meaningless to them. If so, this would explain the results of Experiment 1.

A second issue of Experiment 1 was that the veridical estimates obtained in the noncumulative condition may have been due coincidentally to selecting

populations whose means were separated by just the right amount. If subjects had been evaluating data from a different pair of normal populations, they might not have given such accurate estimates, and in fact subjects might be completely insensitive to differences in pairs of normal populations. Experiment 2 was designed to answer these criticisms.

Method

Design. Three d'(=$(m_1 - m_2)/\sigma$) levels wereused in the experiment: 1.0, 1.6, and 2.2. Experiment 1 used a d' of 1.6. The likelihood ratio (which for our purposes is numerically equal to the odds after seeing one data item) at the mean of the sampled distribution is 1.65, 3.60, and 11.24, respectively, for the three d' levels. The variances for all the populations used in the experiment were the same, and d' was varied by moving the means of the distributions.

The experiment used a within subjects design, with every subject making both aggregated and nonaggregated odds estimates at all three d' levels. Order of presentation of d' levels and the order of making aggregated vs. nonaggregated responses were counterbalanced, but within one d' level subjects made all their responses in one of the aggregation conditions before going to the other aggregation condition. They then proceeded to the next d' level and made both aggregated and nonaggregated estimates in the same order as previously, then continued to the final d' level and did the same thing.

Sequences contained ten sticks each. There were four basic sequences, as follows:

1. Sequence A. This sequence was a random sample from one of the 1.6 d' level populations. Subjects saw this sequence three times during the experiment, once for each d' level. E.g., at some point during the experiment when they were working with d'=2.2, they saw this sequence, but were not told that they might have seen it before, during another d' level.

2. Sequence B. There were three B sequences, one for each d' level. They were constructed by drawing a random sample from a standard normal distribution and converting the sample standard scores to actual samples from the three pairs of populations. For instance, the third stick in sequence B might have a standard score of 1.0, in which case for the small d' level the stick would have 5 3/8" of blue, for the medium d' it would have 5 3/4" of blue, and for the large d' it would have 6 1/8" of blue. Thus all three B sequences were truly repre-

sentative of the populations from which they were drawn.

3. Sequence C. As was the case for B, there were three C sequences, one for each d' level. They were constructed so that the likelihood ratios were the same (or very nearly the same) for all three d' levels. Thus, for instance, if stick #8 had a likelihood ratio of 5.2:1 in favor of the blue population, the stick which would yield that likelihood ratio for the small d' level was located in that position, and similarly for the other two d' levels. Sequence C was a representative sequence from the medium (d'=1.6) level.

4. Sequence D. Three D sequences were constructed, exactly like the C sequences (likelihood ratios constant across d' levels), except that the sequence was a representative sequence from the small d' level.

Twenty sequences were actually constructed: 10 favoring the predominantly blue populations (1 of A and 3 each of B, C, and D) and 10 mirror-image sequences favoring the predominantly yellow populations. All subjects saw all 20 sequences (plus repetitions of sequence A), half of them in a non-aggregating response mode and half in an aggregating mode. For instance, if during the large d' part of the experiment they saw sequence B favoring blue while they were aggregating, then they saw sequence B favoring yellow while doing single-stimulus responses.

The sequences were designed so that there would be a large range of likelihood ratios and of posterior odds. In fact, the veridical likelihood ratios varied from 1:1 to about 365:1, while posterior odds after ten stimuli varied from 12.7:1 to about 13 billion:1.

Subjects. Twenty-four paid male students, run singly or in pairs, served as subjects in the experiment.

Apparatus and Procedure. The apparatus was the same as that used in Experiment 1, except as follows.

Response sheets for Experiment 2 contained six logarithmically spaced scales, representing odds (or likelihood ratios) from 1:1 to 1,000,000:1. There was room at the bottom of the sheet to write in responses greater than 1,000,000:1. Subjects made their estimates by making a mark at the appropriate place on one of the scales or writing in a number if they wanted to make an estimate greater than 1,000,000:1.

Other aspects of instructions and procedures were the same as in Experiment 1, except that only the noncumulative odds and cumulative odds responses were used.

Results. First the data were subjected to a logarithmic transformation, and all analyses were performed on the log transformed responses. As in Experiment 1, the correct hypothesis appears as the numerator of every odds and every likelihood ratio.

Overall, the results were very similar to the results of Experiment 1. A scatterplot with the veridical log odds on the abscissa and the mediam log responses on the ordinate shows that subjects were very accurate when they were not required to aggregate and were conservative when they had to cumulate information. Table 1 shows the results of a regression analysis of such scatterplots. The correlation coefficients in all cases were very high, showing that subjects basically understood how the information in the sticks should affect their estimates. The intercepts were almost zero, indicating little or no bias in favor of one or the other pile. However, the slopes of the lines varied

---
Insert Table 1 about here
---

greatly, depending on whether or not subjects had to aggregate. When they did not aggregate, the slopes were nearly 1.0, indicating that their estimates were extremely accurate, while when they aggregated, the slopes were always less than 1.0, and varied from .329 to .623.

Figure 3 contains scatterplots such as those described above. Figure 3a shows the single-datum estimates as a function of the Bayesian values, while Figure 3b shows the cumulative estimates as a function of the Bayesian values. Of course, the veridical range of the odds greatly exceeded that of the likelihood ratios, but there was a large degree of overlap in the ranges, and it is obvious that over this range the cumulative estimates and the noncumulative estimates did not lie along the same function, even within the 10:1 range. Furthermore, it is apparent that both sets of points are linear, not S-shaped.

Insert Figure 3 about here

DuCharme (1970) used sequences of variable length, ranging from one to seven stimuli long. Thus when his subjects saw the first trial of a sequence they never knew whether that would be the only trial or whether there would be

other trials following it. Therefore his noncumulative estimates were always trial #1 responses. In the current experiment, if subjects were in the aggregating condition they always knew that the first stick in a sequence would be followed by nine others, while in the nonaggregating condition they knew that each stick was completely independent of all others. Thus one can look at two kinds of noncumulative estimates: all the nonaggregation responses, and the trial #1 responses in the aggregation condition. A regression analysis was performed for the Bayesian log likelihood ratio vs. the median subject log likelihood ratios for each trial number. As the misaggregation hypothesis predicts, in the nonaggregating conditions subjects' responses were veridical for all trial numbers, since they were effectively all first responses. The slopes were all very close to 1.0. However, when subjects cumulated the story was quite different. Since trial #1 is theoretically identical to the nonaggregation condition, trial #1 responses should have been identical to the nonaggregation responses, while trial #2 and all ensuing trials should have shown conservatism. The slope of the best fitting line for trial #1 responses was considerably less than 1.0; it was .59. Thus, although trial #1 responses were more veridical than the other cumulative responses (their slopes were about .38), they were not identical to the nonaggregated responses. This means that trial #1 responses, if plotted on Figure 3, would lie somewhere between the nonaggregated responses and the aggregated responses. DuCharme's trial #1 responses fell along the same curve as the cumulated estimates.

An implication of the response bias hypothesis is that within a subject's veridical range, he will never show conservatism, no matter how much aggregation he must perform. Figure 4 speaks to this question. It plots the cumulative log

---

Insert Figure 4 about here

---

odds for certain sequences as a function of stimulus number. Figures 4a and 4b show conservatism when subjects aggregated but display nearly veridical cumulative log odds as inferred from their noncumulative estimtes. Since the final odds were well outside the 10:1 range this result was not very surprising. However, Figures 4c and 4d are for sequences in which the odds never exceeded 10:1 until the tenth stimulus, and even then were less than 15:1; here also there

was obviously conservatism in the estimates when subjects aggregated. There are 24 figures such as the four displayed in Figure 4 (one for each sequence); some show slightly more conservatism than those in Figure 4, some show less. However, in all cases the overall conclusion is the same: when subjects must aggregate information, they will be conservative. When they must judge only a single item of information, they are generally veridical, or perhaps even a little radical.

The second question with which the experiment was concerned was whether or not subjects are sensitive to differences in d'. DuCharme's (1970) study also looked at this question; his results indicated that they are. Table 2 shows that subjects in the current experiment were sensitive to d' also. Table 2 shows the

---
Insert Table 2 about here
---

d' inferred from subject's median log estimate taken at the mean of the distributions. Thus, for instance, if subjects correctly perceived the distributions in which d'=1.6, they would give a likelihood ratio estimate of 3.60 when they saw a stick whose length of blue was the same as the mean length of the distribution. When subjects did not cumulate, the inferred d' is almost identical with the actual d'; when they aggregated, they responded as though d' were smaller than it actually was. However, in either the cumulative or noncumulative condition, they perceived that the three pairs of populations were really different, and responded in the correct direction.

Recall that the sequences of stimuli were constructed to have carefully chosen properties. One set of sequences (A) involved showing the same actual sticks at three times during the experiment, once for each d' level. If subjects had not been sensitive to d', they would have given approximately the same estimates for those sticks all three times the sequence was displayed, regardless of the d' they were dealing with at the time. An analysis of the regression lines of log likelihood ratios as a function of length of blue (for sequence A) showed that the slopes of the lines for the three d' levels were significantly different from one another and were virtually identical to the theoretically correct slopes. ($F = 14.17$, with 2 and 3 d.f.; $p < .05$.) The subjects in the experiment were sensitive to the differences in d', and responded appropriately.

Sequence set B involved displaying sticks whose standard scores were constant across d' levels. It is possible that subjects might assign the same likelihood ratios to certain standard scores no matter what the d' is. A regression analysis of median log likelihood ratio estimates for sequence B as a function of the standard scores of the stimuli again verified the subjects' ability to respond correctly to d'. Again the slopes of the lines for the three d' levels were significantly different from one another and were very nearly the same as the theoretically correct slopes. ($F = 22.31$ with 2 and 3 d.f.; $p < .025$.)

The third and fourth sets of sequences (C and D) were constructed so that the likelihood ratios remained constant across the three d' levels. If subjects responded correctly to d', they should have given approximately the same response to a given stimulus number (e.g., stick #4 in sequence D) no matter which d' level they were working with. Regression analyses showed that the slopes of the lines of median log likelihood ratios as a function of Bayesian log likelihood ratios were not significantly different from one another nor from the theoretically correct responses. ($F = .962$ with 2 and 3 d.f.)

All analyses discussed so far have been with median responses. How well do these results describe any one individual? To answer that scatterplots and regression analyses were obtained for each subject. The results were surprisingly orderly. All correlation coefficients comparing Bayes with the subject's responses were between .46 and .97, with a median of .89 for noncumulative responses and .87 for cumulative responses. The slopes of the best fitting lines for noncumulative responses varied from .08 to 3.81 and for the cumulative estimates from .06 to 1.00. Those medians were 1.18 and .41 respectively. There was a great deal of variability in the slopes of the lines, as one would expect, but all subjects were well fit by a straight line, indicating orderliness in their responses.

Discussion. The results of the experiment clearly support the misaggregation hypothesis. In every analysis subjects' responses when they were aggregating were different from their responses when they did not have to aggregate. Two things about these results are rather puzzling. The first is the fact that first-trial responses in the aggregation condition were not identical to non-aggregated responses, although theoretically they should be. The second is that

these results were strongly at variance with the results of DuCharme's (1970) experiments.

Both the DuCharme experiments and Experiment 2 used normal data generators, and both used several levels of diagnosticity (d'). In both cases it was apparent that subjects responded correctly to diagnosticity, which is an important and gratifying result. However, DuCharme's results supported the response bias hypothesis (although not in its simplest form). Why did his results strongly support one hypothesis while the results of the current experiment strongly support a different hypothesis? There were some important differences in the two experiments: types of stimuli (numerical vs. physical), length of sequences (short vs. long), knowledge of the length of the sequences (unknown and variable vs. known and fixed), population display (none or orderly histograms vs. random sample), and of course, the experimenters themselves. Experiment 3 looks at some of these differences in an attempt to explain why DuCharme's results were so different from those of Experiment 2.

Experiment 3

Experiment 3 varied sequence characteristics and population displays.

In the DuCharme experiments, the sequence lengths were variable (1-7 stimuli in one study, 1-5 in the other), and the subjects did not know how long a sequence would be until they reached its final trial. In Experiments 1 and 2 of this paper, if subjects were in a noncumulative response condition they always saw independent stimuli; if they were in a cumulative response mode they knew that each sequence would be 10 stimuli long. In Experiment 3 each subject received all three types of sequences: variable, fixed, and independent trials.

In the first experiment of the DuCharme study, he used no population display at all, but he was dealing with familiar populations (heights of men and women). It seemed impractical to use no display when dealing with pickup sticks as stimuli, so no such condition was employed in Experiment 3. In DuCharme's second experiment he used orderly histograms of 100 samples from each population. Experiments 1 and 2 in the current series used randomly ordered samples of about 100 (97 to be exact) as the display, with the mean of the distribution clearly marked on the display. DuCharme and Peterson (1968) had used

overlapping normal density curves to display their populations. These three types of displays (histogram, randomly ordered sample, and overlapping density curves) were used in Experiment 3.

Method

Design. The data generating populations were normally distributed. Only one d' level was used: 1.6, which yields a likelihood ratio of 3.60 at the mean of the distribution.

There were three display conditions: random samples (used in Experiments 1 and 2), histograms, and overlapping normal curves. DuCharme and Peterson (1968) had used overlapping normal curves in their initial heights-of-men-and-women experiment, and DuCharme (1970) used the histogram in his dissertation experiments.

Subjects responded to three types of sequences: single sticks (comparable to the noncumulative conditions of Experiments 1 and 2), fixed length sequences each containing six sticks (comparable to the cumulative conditions of Experiments 1 and 2), and variable length sequences containing from one to seven sticks (comparable to DuCharme's experiment). Each subject made judgments of all three kinds of sequences, but there were four different orders of presentation of the sequences.

The single-stick (noncumulative) part of the experiment contained 27 sticks, ranging from solid yellow to solid blue, including one that was half blue and half yellow.

There were 12 fixed-length sequences of six sticks each. Six of the sequences were from the predominantly blue population while the other six were mirror-image sequences favoring the predominantly yellow population.

The variable-length sequences consisted of 4 sequences each of lengths two through seven, plus 16 single sticks. There were 2 sequences favoring blue and 2 mirror-image sequences favoring yellow for each sequence length. The 16 sticks were randomly inserted among the longer sequences throughout the variable condition.

Subjects. Forty-eight male students served as subjects in the experiment; they were equally divided among the three display conditions and four sequence presentation order conditions. They were run in groups of one, two, or three.

Apparatus. All apparatus and response sheets were like those in Experiment 2, with the addition of the histograms and overlapping normal curves used in those two display conditions.

Procedure. The procedure was identical to that of Experiment 2, except that instructions for variable sequences were inserted at the apporpriate place during the experiment. In the variable part of the experiment, subjects were instructed to assume that the experimenter had selected one of the populations at random, then had drawn some unspecified number of sticks from it. They would see the sticks in the order in which they had been drawn, and were to revise their odds after each new stick was presented, continuing this until she informed them that there were no more sticks in that sequence. Subjects were never told that seven was the maximum number of sticks in a sequence.

Let F stand for the fixed-length sequences, V for the variable-length sequences, and S for the single-sticks condition. Then the four sequence presentation orders were: FVS, VFS, SFV, and SVF.

Results

First the data were subjected to a logarithmic transformation, and all analyses were performed on the log transformed responses. As before, the correct hypothesis was in the numerator of all odds and likelihood ratios.

Scatterplots and regression analyses of veridical log odds vs. median subject log responses were obtained for all conditions. The display condition had little or no effect on responses. The correlation coefficients were high (greater than .96) in all cases; the slopes of the lines varied depending on sequence type and trial number. The slopes in the histogram condition were slightly lower than the slopes for the other conditions, but not significantly so. ($F = .14$, with 2 and 18 d.f.)

The data told much the same story for the different orders of presentation. Again the correlation coefficients were high (greater than .95 in all cases) and the slopes of the lines varied depending on sequence type and trial number. There were no significant differences between presentation orders. ($F = 1.54$, with 3 and 24 d.f.)

Since there were no differences between the various order and display conditions, scatterplots and regression analyses were obtained collapsing over all conditions, using the median for all 48 subjects. Table 3 summarizes these results. When subjects were in a nonaggregating (single stick) condition, they were virtually veridical, as indicated by the slope of 1.022. When they were

---

Insert Table 3 about here

---

in an aggregation condition, with either fixed or variable length sequences, they were conservative, and this conservatism showed up on Trial 1 as well as on later trials. However, they were less conservative on Trial 1 than on later trials. This result agrees with the result of Experiment 2 in which the slope of the line on Trial 1 of the aggregating condition lay somewhere between the veridical nonaggregated responses and the more conservative later-trial aggregated responses.

DuCharme (1970) displayed the scatterplot of his data but did not report regression values. A regression analysis of his Figure 1 data yielded some interesting results. Table 4 compares Experiment 2, Experiment 3, and DuCharme's results. DuCharme did not have a condition comparable to the single-

---
Insert Table 4 about here
---

stick condition; his results were all based on a variable-length condition. His results correspond favorably with those of the present experiments, with the Trial 1 responses being somewhat more veridical than later-trial responses, and both being less veridical than nonaggregation condition responses.

Figure 5 shows the scatterplots of the median subject log odds vs. the Bayesian log odds for Trial 1 in the three sequence conditions. There is no apparent difference between the fixed sequences and the variable sequences. There does appear to be a slight S shape in the plot for single sticks. Note,

---
Insert Figure 5 about here
---

however, that it straightens out near log odds of $\pm$ 2:1. It may well be that the S shape would not have been so noticeable had there been more stimuli whose veridical log odds were greater than 1.5:1. Unfortunately there are only a few data points in that range, and one cannot tell whether the subjects would continue to flatten their curves or would straighten them out if more intermediate stimuli were presented. The variable and fixed plots are obviously linear.

Cumulative values are displayed in Figure 6. Once again the data exhibit linearity over a wide range of veridical values, and this is true for the

variable sequences as well as the fixed sequences.

Insert Figure 6 about here

As was done in Experiment 2, scatterplots and regression analyses were obtained for individual subjects. Once again individual subject's estimates were very orderly, and showed the same properties found in the median analyses.

These results, while gratifying to a proponent of the misaggregation hypothesis, still leave unsettled the question of why they differed so strongly from DuCharme's results. Close inspection of his data yields a clue to the puzzle. In his first experiment, 36% of the Bayesian odds were larger than 100:1, but only 2.8% of his median odds were at least that large. In Experiment 2 of the current paper, 42% of the Bayesian odds were greater than 100:1 and 6.3% of the median estimates were at least that large. Looked at another way, DuCharme's subjects made estimates greater than 100:1 in only 7.7% of the instances when a Bayesian would have done so. The subjects in Experiment 2 made estimates that large or larger in 15% of the instances when it was appropriate to do so. Apparently DuCharme's subjects were considerably more reluctant to estimate larger values than the subjects in Experiment 2. Why? One reason might be that his response device encouraged subjects to put an upper bound on their responses.

The device consisted of six logarithmically spaced odds scales, only one of which was visible at any time. Subjects set a sliding lever at the odds they selected, then wrote down the odds on a sheet of paper. For the next trial they moved the lever from its previous position to a new odds location. The six scales went from 1:1 to 1,000,000:1, with the first one going from 1:1 to 10:1, the second from 10:1 to 100:1, and so on. In order to go from one scale to the next, DuCharme's subjects had to physically rotate the metal bar to which the scales were attached. Thus an extra physical action was necessary to make any response greater than 10:1; estimates larger than 100:1 required two rotations of the scale. More evidence that the response device might be encouraging flattened response curves appears from studying prior odds data. DuCharme in one part of his study varied prior odds, using values

of 2:1, 5:1, 10:1, and 100:1. If one looks at the number of times that a Bayesian would have rotated the scale at least twice from where it was originally set, he finds that although this should occur 16 times, there were no instances of double rotation in the median estimates. A similar look at DuCharme's second experiment reveals that a Bayesian would rotate the scale twice 74 times, but his median subject did so only once. So it may be that a considerable portion of the S shape reported by DuCharme is an artifact revolving around a response device that encourages subjects to set an upper bound on their estimates.

Experiments 2 and 3 reported here used a page with logarithmically spaced odds scales printed on it. Thus the entire range available to DuCharm's subjects was visible at all times to the subjects in these experiments. In addition to the figures already reported for Experiment 2, in Experiment 3 there were 62 instances when the Bayesian values were greater than 100:1. The median subjects' estimates were at least 100:1 in 33 of those instances.

Discussion

Many experimenters have studied conservatism in human inference tasks during the past ten years or so. Some early experiments supported the misperception hypothesis, but more recent experiments favored either the misaggregation or the response bias explanations. The results of the studies reported in this paper strongly support the notion that misaggregation is the predominant explanation for conservative inference. These findings, however, conflict with those of DuCharme (1970) and to a lesser extent with those of DuCharme and Peterson (1968). There is evidence that at least some of the response bias results are artifactual. Any serious student of the field must, however, admit that in all likelihood misperception, misaggregation, and response biases all contribute to conservatism. The real questions of importance then become finding the manner in which each phenomenon contributes to conservatism and the best way of avoiding or compensating for this nonoptimal behavior.

Because of the rather extensive research literature that has developed around the questions of when and why people make inferences conservatively, the appropriateness of those questions as topics for research has been questioned recently. A referee of an earlier version of this paper said "... it has be-

come rather widely accepted in the literature that the question of the causes of conservatism is not a fruitful psychological question. The phenomenon of conservatism is as much a property of the method of data analysis as it is of the subject's responses (Anderson and Shanteau, 1970; Pitz, 1970; Rapaport and Wallsten, 1972; Shanteau, 1970, 1972; Winkler and Murphy, 1973)." That list of references could be brought up to date by adding Slovic (1972) and Hogarth (1975) to it.

Review of the critics' complaints brings out three main themes:

1. Since Bayes's Theorem is a wholly artificial external standard having nothing to do with known behavior, it is inappropriate to compare human inferences with it.

2. The phenomena of conservatism are both subject- and task-dependent. In particular, studies in realistic settings with professional inference-makers as subjects do not show the phenomenon. Consequently, it is an inconsequential laboratory artifact.

3. Research that compares human inference-making with Bayes's Theorem implicitly or explicitly assumes that human inference is, to some extent, Bayesian in character. It is not, and consequently the research is irrelevant.

Our answers to these complaints go as follows:

1. We just don't understand this complaint at all. Research on formal inference uses the syllogism as criterion against which to compare behavior. Research on grammar uses the rules of correct grammar for comparison. Research on mental arithmetic uses the rules of arithmetic. Why should not research on human probabilistic inference compare such inferences with the output of the formally correct rules for making them? If correctness is to be ruled out of use by psychologists in search of dependent variables, what is to become, for example, of most research on verbal learning or choice reaction time?

2. Conservatism is indeed both subject- and task-dependent. So far as we know, so are all other phenomena of human intellectual behavior. Conservatism is certainly sufficiently pervasive, over both people and tasks, to deserve study and explanation. And while some professional inference-makers seem not to be conservative (e.g., Winkler, 1971 and Peterson, Snapper, and Murphy, 1972), others clearly are (e.g., Kelly and Peterson, 1970 and Zlotnick, 1968). It probably depends on detailed characteristics of the inference tasks. We have

speculations about those task characteristics that do and do not favor conservatism, but need not review them here. For an extensive discussion, see Goodman (1973).

3. We regard human inferences as being described, to some extent and as a first approximation, by Bayes's Theorem, and cite, for example, Figures 1 and 4 of this paper in support of that belief. Critics who feel otherwise often cite Tversky and Kahneman's (1974) finding that the prior probability is ignored after the first datum is presented as evidence to the contrary. Their phenomenon too is task-dependent; Peterson and DuCharme (1967) found the opposite. Detailed research on actual effects of task characteristics, such as is reported here, seems to us more likely to clarify this kind of problem then is sweeping assertions that such details are irrelevant. The question "are men Bayesian?" is obviously silly; the answer is no. The right questions are: what kind of first approximation to human inference-making is offered by Bayes's Theorem in what situations; and how can that first approximation be improved on?

References

Anderson, N.H. & Shanteau, J.C. Information integration in risky decision making. Journal of Experimental Psychology, 1970, 84, 441-451.

DuCharme, W.M. A review and analysis of the phenomenon of conservatism in human inference. Applied Mathematics and Systems Theory Technical Report No. 46-5, December 30, 1969, Rice University.

DuCharme, W.M. Response bias explanation of conservative human inference. Journal of Experimental Psychology, 1970, 85, 66-84.

DuCharme, W.M. & Peterson, C.R. Intuitive inference about normally distributed populations. Journal of Experimental Psychology, 1968, 78, 269-275.

Edwards, W. Conservatism in human information processing. In Kleinmuntz, B. (Ed.), Formal Representation of Human Judgment. New York: Wiley, 1968.

Edwards, W., Phillips, L.D., Hays, W.L., & Goodman, B.C. Probabilistic information processing systems: Design and evaluation. IEEE Transactions on System Science and Cybernetics, 1968, SSC-4, 248-265.

Goodman, B.C. "Direct Estimation Procedures for Eliciting Judgments about Uncertain Events," Technical Report. University of Michigan Engineering Psychology Laboratory, November 1973.

Hogarth, Robin M. Cognitive processes and the assessment of subjective probability distributions. Journal of the American Statistical Association, 1975, 70, 271-289.

Kelly, C.W., III & Peterson, C.R. Probability estimates and probabilistic procedures in current-intelligence analysis. Report on Phase I, June 1970-December 1970, FSC 71-5047, Federal Systems Division, International Business Machines Corporation, Gaithersburg, Maryland.

Peterson, C.R. & DuCharme, W.M. A primacy effect in subjective probability revision. Journal of Experimental Psychology, 1967, 73, 61-65.

Peterson, C.R., DuCharme, W.M. & Edwards, W. Sampling distributions and probability revisions. Journal of Experimental Psychology, 1968, 76, 236-243.

Peterson, C.R., Snapper, E.J. & Murphy, A.H. Credible interval temperature forecasts. Bulletin of the American Meteorological Society, October 1972, 53. 966-970.

Pitz, G.F. On the processing of information: Probabilistic and otherwise. Acta Psychologica, December 1970, 34, 201-213.

Rapaport, A., & Wallsten, T.S. Individual decision behavior. Annual Review of Psychology, (1972), 23, 131-176.

Shanteau, J.C. An additive model for sequential decision making. Journal of Experimental Psychology, 1970, 85, 181-191.

Shanteau, J.C. Descriptive versus normative models of sequential inference judgment. Journal of Experimental Psychology, 1972, 93, 63-68.

Slovic, P. From Shakespeare to Simon: Speculations--and some evidence--about man's ability to process information. Oregon Research Institute Monograph, 1972, Vol. 12, No. 2.

Slovic P. & Lichtenstein, S. Comparison of Bayesian and regression approaches to study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, September 27, 1974, 185, 1124-1131.

Wheeler, G., & Beach, L.R. Subjective sampling distributions and conservatism. Organizational Behavior and Human Performance, 1968, 3, 36-46.

Winkler, R.L. Probabilistic prediction: Some experimental results. Journal of the American Statistical Association, December 1971, 66, 675-685.

Winkler, R.L. & Murphy, A.H. Experiments in the laboratory and the real world. Organizational Behavior and Human Performance, 1973, 10, 252-270.

Zlotnick, J. A theorem for prediction. Foreign Service Journal, 1968, 45(8), 20.

Footnotes

This research was supported in part by the National Aeronautics and Space Administration under Grant NGL-23-005-171 to the Engineering Psychology Laboratory, University of Michigan, monitored by the Ames Research Center, National Aeronautics and Space Administration, and in part by the Advanced Research Projects Agency of the Department of Defense and was monitored by ONR under Contracts N00014-67-A-0181-0049 and N00014-75-C-0487.

Requests for reprints should be sent to Ward Edwards, Social Science Research Institute, University of Southern California, Los Angeles, California, 90007.

TABLE 1

Experiment 2: Regression Analyses of Median Subject Log Estimates as a Function of Bayesian Log Odds

d' level	Noncumulative Responses			Cumulative Responses		
	Correlation Coefficient	Slope	Intercept	Correlation Coefficient	Slope	Intercept
All	.963	1.118	-.022	.942	.381	-.051
1.0	.967	1.514	-.015	.949	.623	.046
1.6	.978	.998	-.016	.982	.385	-.033
2.2	.978	1.068	-.033	.985	.329	-.247

TABLE 2

Experiment 2: d' Inferred from Subjects' Responses

True 'd	Inferred 'd		Likelihood Ratio of an Observation Taken at the Mean		
	From Noncumulative Responses	From Cumulative Responses	True Value	From Noncumulative Responses	Inferred from Cumulative Responses
1.00	1.23	.76	1.65	2.14	1.33
1.60	1.60	.95	3.60	3.59	1.58
2.20	2.27	1.17	11.24	13.25	1.98

TABLE 3

Experiment 3: Regression Analyses of Median Subject Log Estimates as a Function of Bayesian Log Odds

Trial Number	Single Sticks		Variable Sequences		Fixed Sequences	
	Correlation Coefficient	Slope	Correlation Coefficient	Slope	Correlation Coefficient	Slope
1	.978	1.022	.961	.766	.996	.757
2-7			.989	.590	.991	.574
All	.978	1.022	.984	.601	.990	.579

TABLE 4

Summary of Refression Analyses for Median Subject Log Estimates as a Function of Bayesian Log Odds for Experiment 2, Experiment 3, and DuCharme (1970)

Experiment	Independent Trials		Trial 1		Trials 2-10	
	Correlation Coefficient	Slope	Correlation Coefficient	Slope	Correlation Coefficient	Slope
Experiment 2	.939	.909	.912	.590	.947	.399
Experiment 3	.978	1.022	.961	.766	.989	.590
DuCharme (1970)			.954	.515	.977	.324

Figure Captions

Figure 1. Experiment 1: mean inferred log odds as a function of Bayesian log odds for four response modes. Circles indicate two or more data points at the same coordinates. The solid line represents perfect Bayesian performance; the dashed lines are regression lines fitted to the data points.

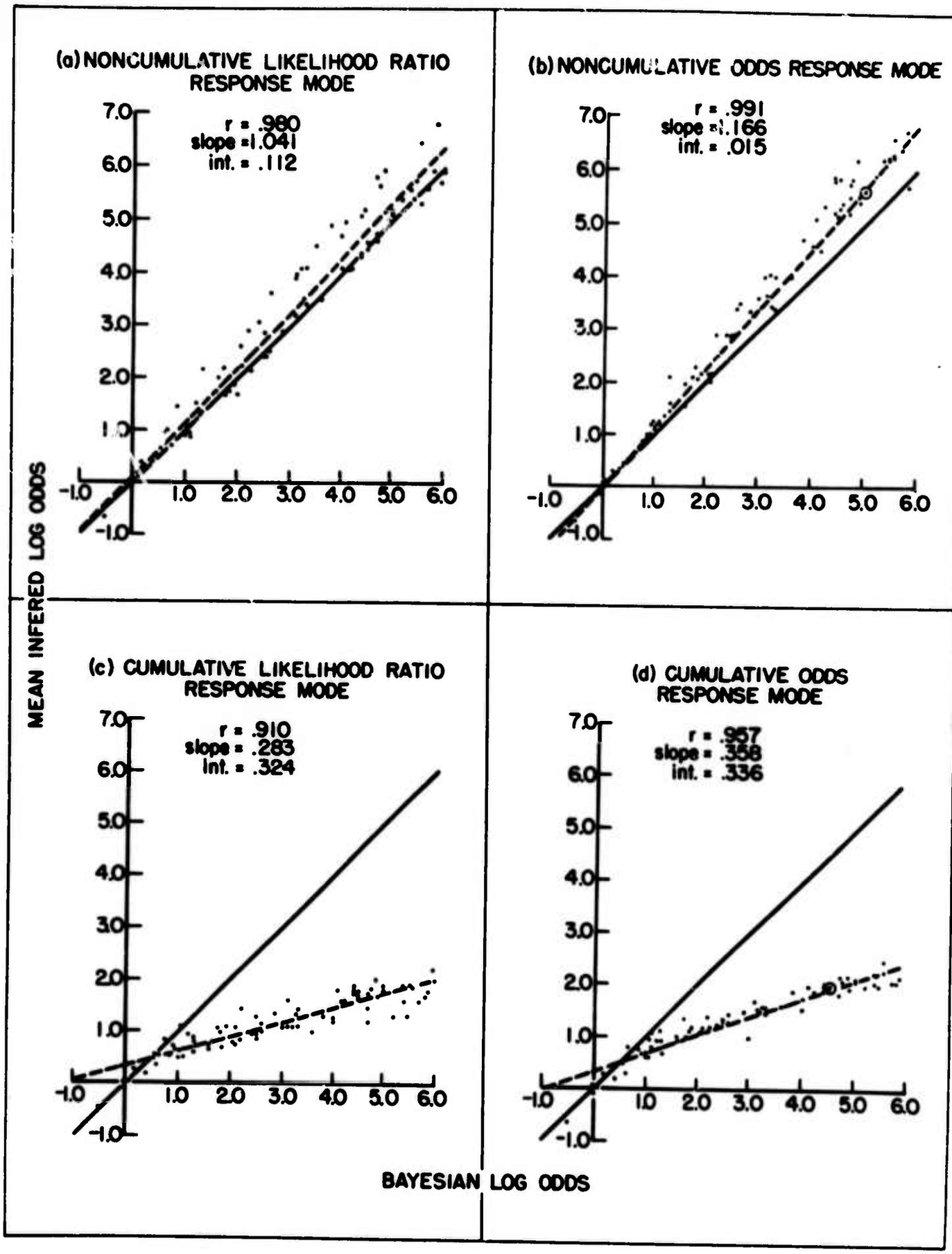
Figure 2. Experiment 1: slope of the best-fitting regression lines for mean inferred log likelihood ratios as a function of Bayesian log likelihood ratios, by trial number. Slopes are plotted on a logarithmic scale, which equalizes the effect of both extreme and conservative deviations from optimal.

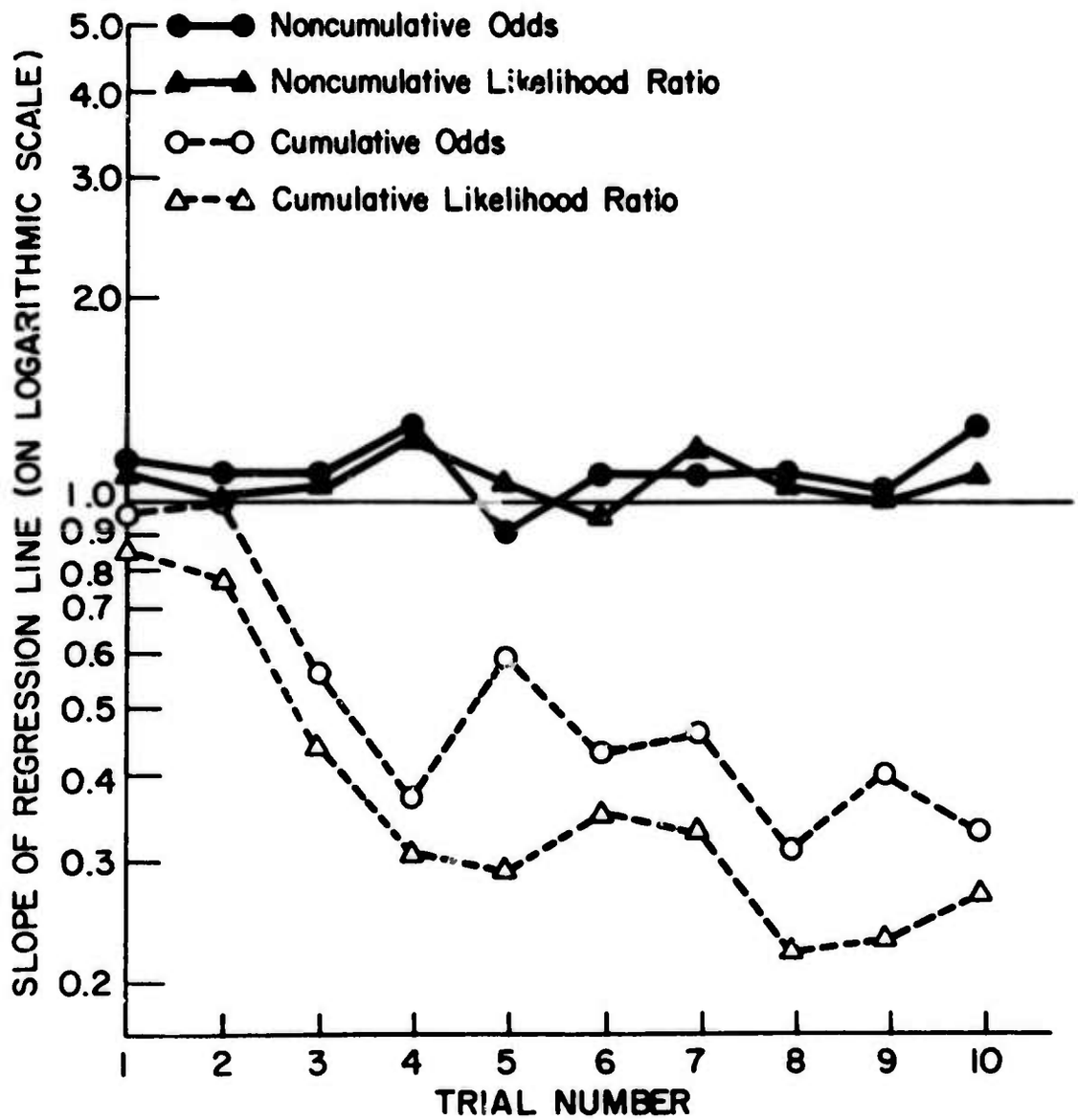
Figure 3. Experiment 2: median estimated log odds as a function of Bayesian log odds for independent trials and cumulative trials. Circles represent two or more data points at the same coordinates.

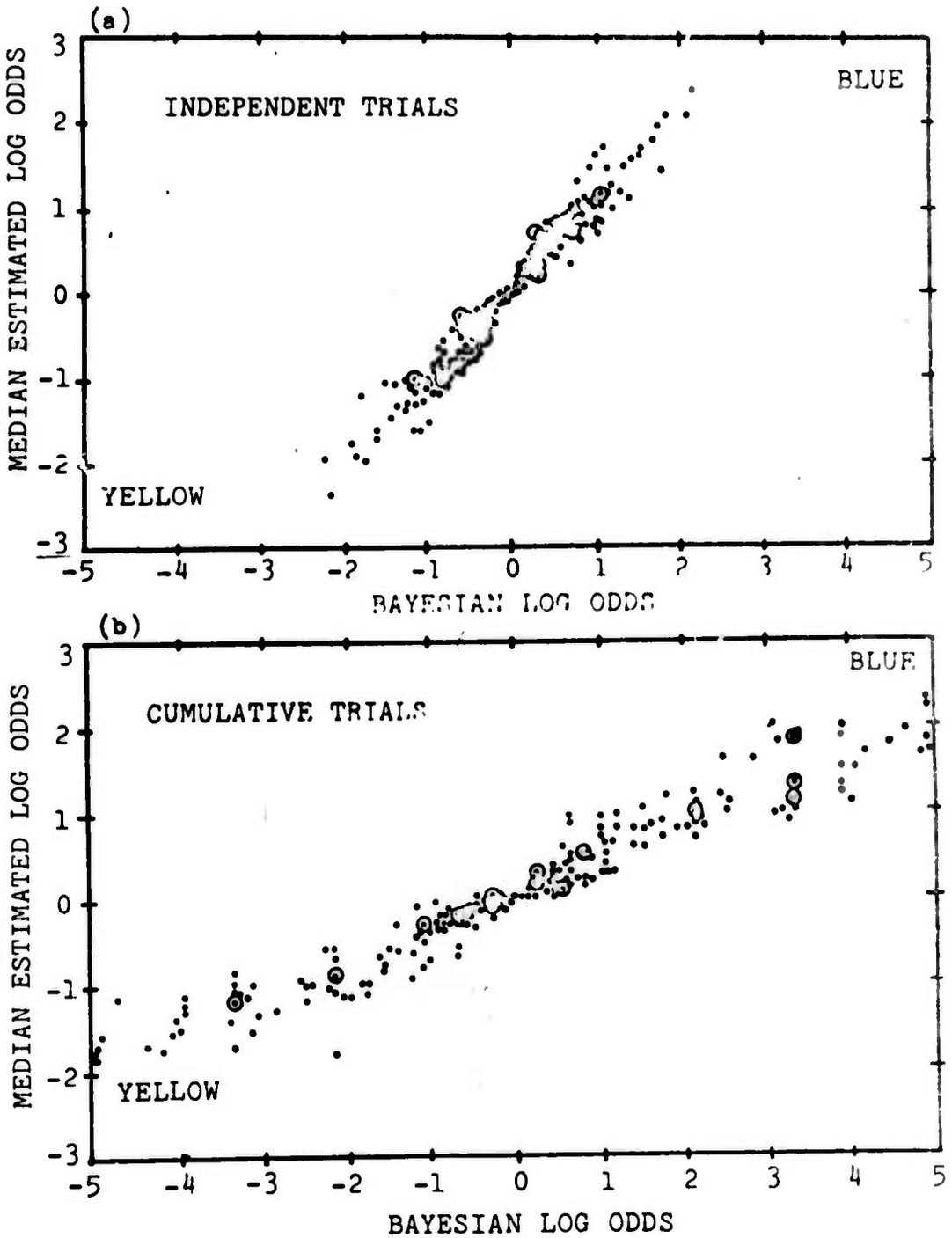
Figure 4. Experiment 2: median estimated log odds for selected sequences as a function of trial number.

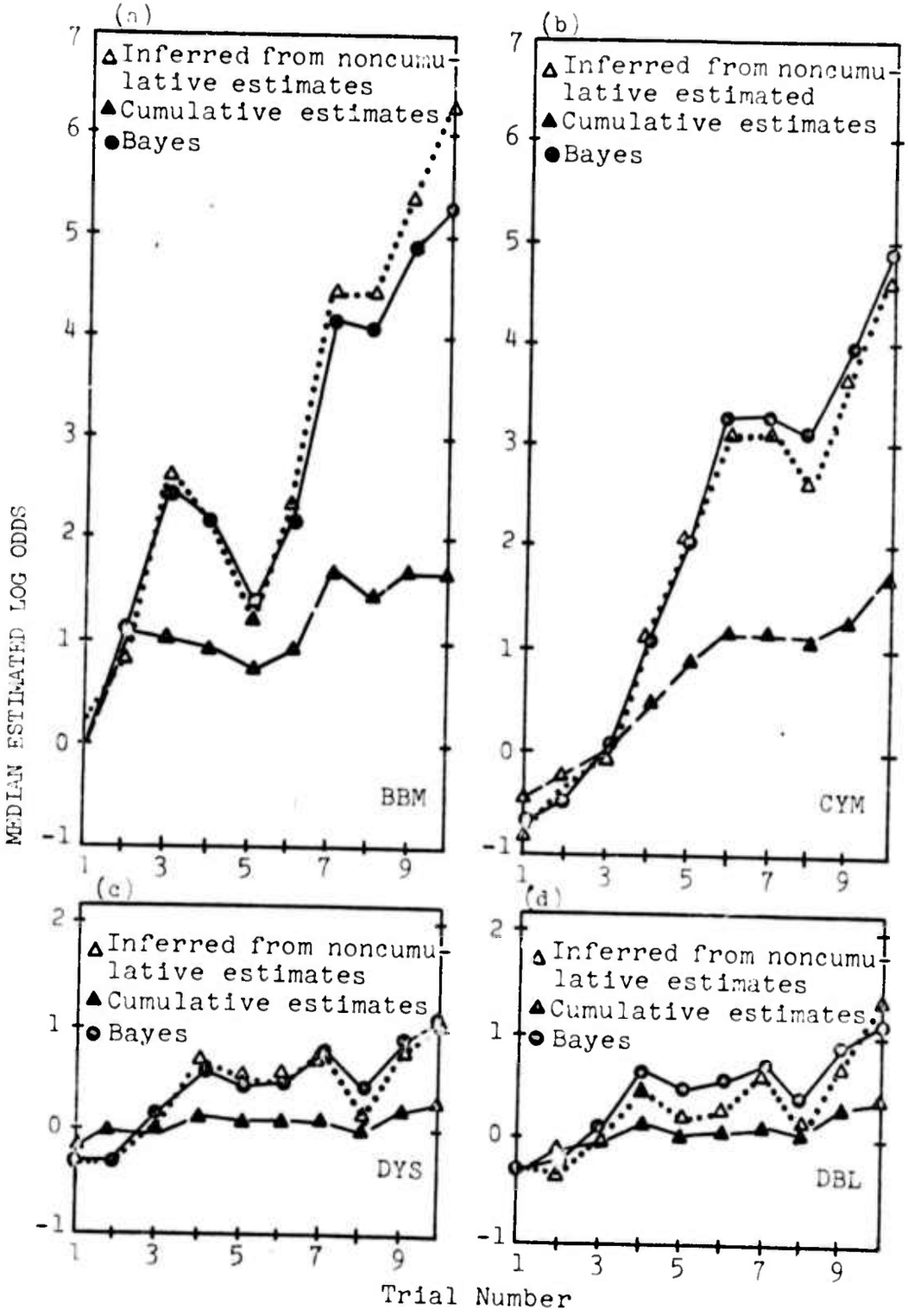
Figure 5. Experiment 3: median estimated log odds as a function of Bayesian log odds for Trial 1 in three sequence conditions.

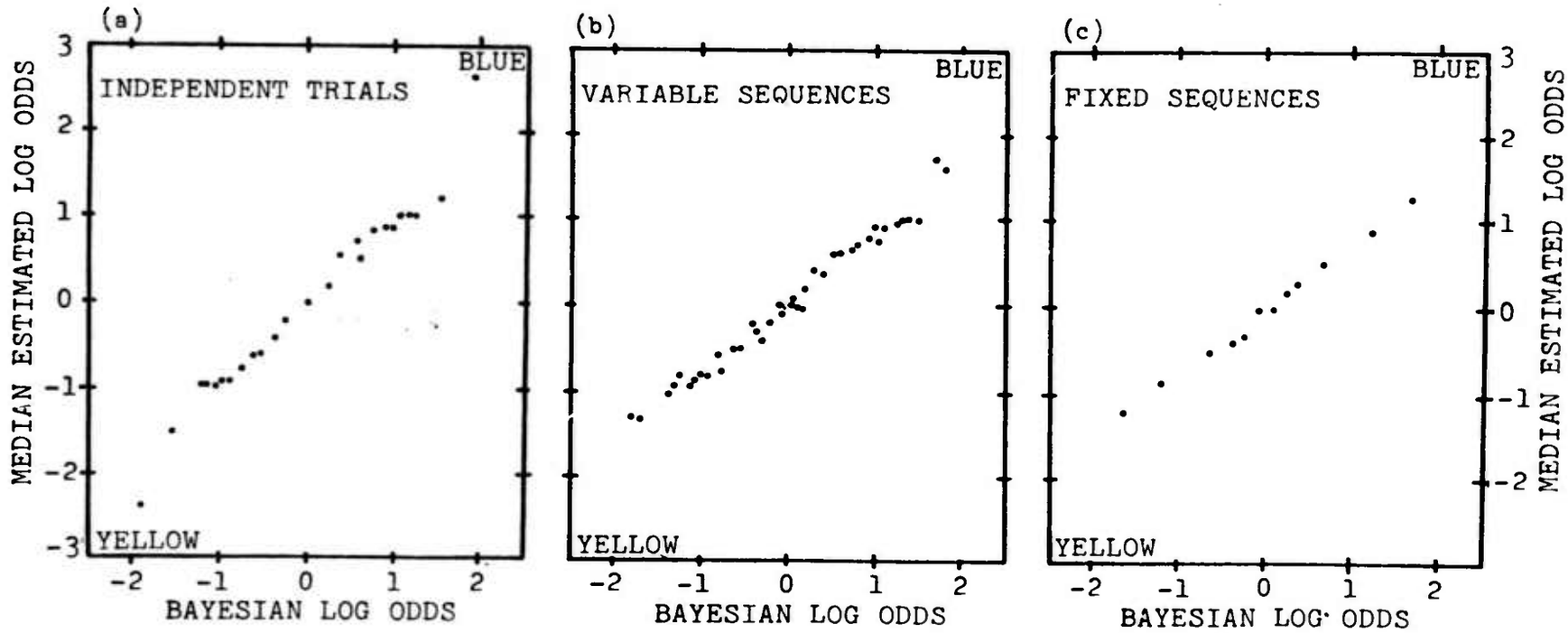
Figure 6. Experiment 3: median estimated log odds as a function of Bayesian log odds for variable sequences and fixed sequences. Circles represent two or more data points at the same coordinates.

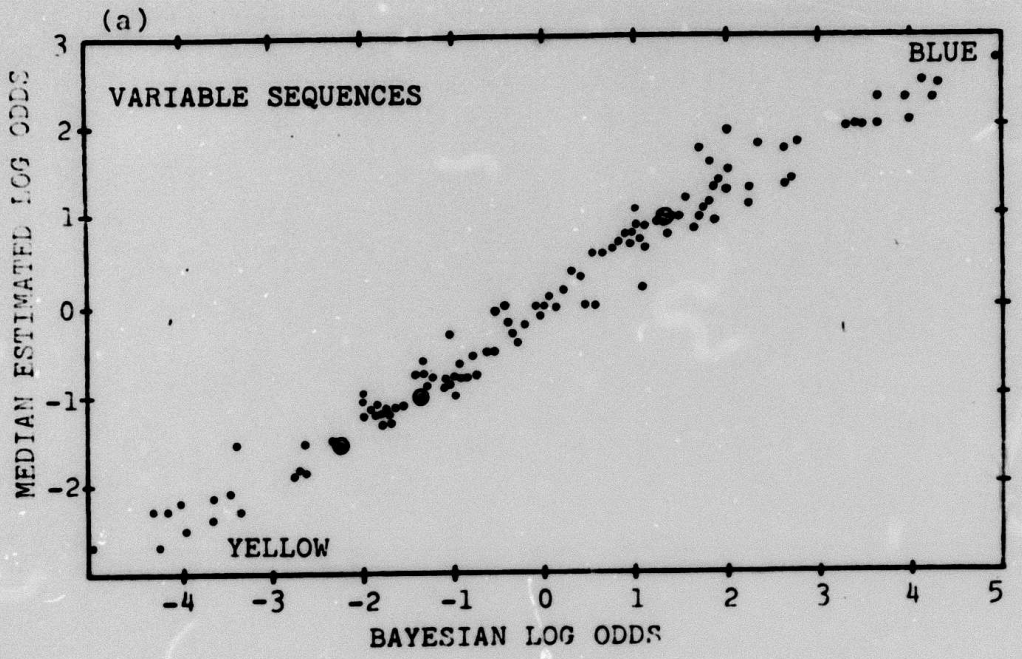

(a) NONCUMULATIVE LIKELIHOOD RATIO RESPONSE MODE
r = .980
slope = 1.041
int. = .112
(b) NONCUMULATIVE ODDS RESPONSE MODE
r = .991
slope = 1.166
int. = .015
(c) CUMULATIVE LIKELIHOOD RATIO RESPONSE MODE
r = .910
slope = .283
int. = .324
(d) CUMULATIVE ODDS RESPONSE MODE
r = .957
slope = .358
int. = .336
MEAN INFERED LOG ODDS
BAYESIAN LOG ODDS
7.0
6.0
5.0
4.0
3.0
2.0
1.0
-1.0
1.0
2.0
3.0
4.0
5.0
6.0

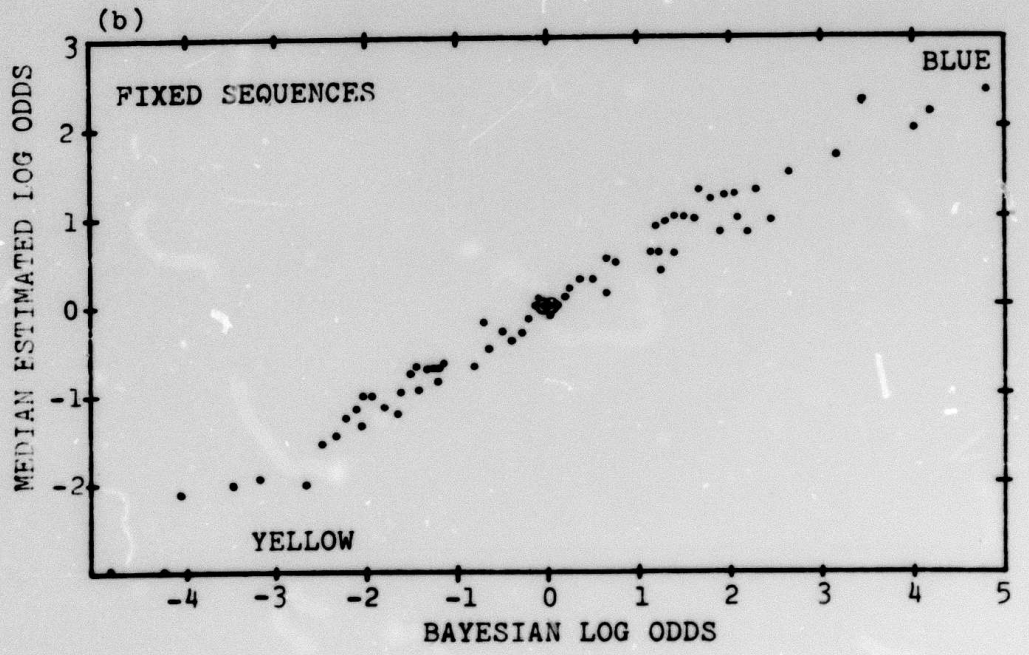

Noncumulative Odds
Noncumulative Likelihood Ratio
Cumulative Odds
Cumulative Likelihood Ratio
SLOPE OF REGRESSION LINE (ON LOGARITHMIC SCALE)
5.0
4.0
3.0
2.0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1
2
3
4
5
6
7
8
9
10
TRIAL NUMBER


(a)
INDEPENDENT TRIALS
BLUE
YELLOW
MEDIAN ESTIMATED LOG ODDS
BAYESIAN LOG ODDS
(b)
CUMULATIVE TRIALS
BLUE
YELLOW
MEDIAN ESTIMATED LOG ODDS
BAYESIAN LOG ODDS


(a)
Inferred from noncumulative estimates
Cumulative estimates
Bayes
BBM
(b)
Inferred from noncumulative estimated
Cumulative estimates
Bayes
CYM
(c)
Inferred from noncumulative estimates
Cumulative estimates
Bayes
DYS
(d)
Inferred from noncumulative estimates
Cumulative estimates
Bayes
DBL
MEDIAN ESTIMATED LOG ODDS
Trial Number


(a)
INDEPENDENT TRIALS
BLUE
YELLOW
MEDIAN ESTIMATED LOG ODDS
BAYESIAN LOG ODDS
(b)
VARIABLE SEQUENCES
BLUE
YELLOW
BAYESIAN LOG ODDS
(c)
FIXED SEQUENCES
BLUE
YELLOW
BAYESIAN LOG ODDS
MEDIAN ESTIMATED LOG ODDS
3
2
1
0
-1
-2
-3


(a)
VARIABLE SEQUENCES
BLUE
YELLOW
MEDIAN ESTIMATED LOG ODDS
BAYESIAN LOG ODDS
3
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
5


(b)
FIXED SEQUENCES
BLUE
YELLOW
MEDIAN ESTIMATED LOG ODDS
BAYESIAN LOG ODDS
3
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
5